Identifying chemical entities in patents using brown clustering and semantic similarity

Andre Lamurias*[1], Manuel Lobo[1], Marta Antunes[1], Luka A. Clarke[2], and Francisco M. Couto[1]

[1] LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] BioISI: Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract. This paper presents the system we developed for the CHEMDNER task of BioCreative V. This system was adapted from the IICE framework, which combines Conditional Random Fields, implemented by Stanford NER, brown clustering, implemented by Percy Liang's C-based algorithm and a semantic similarity based on the h-index concept, applied to the ChEBI ontology. For the CEMP subtask, we obtained a maximum precision of 86.981% (Run 4), maximum recall of 92.327% (Run 3) and maximum F-measure of 84.656% (Run 2), while for the CPD subtask, we obtained maximum specificity of 94.271% (Run 4), maximum sensitivity of 97.875% (Run 3) and maximum MCC of 80.254% (Run 2). Our system achieved the best specificity and third best recall of the task.

Key words: CRFs, Random Forests, Semantic Similarity, ChEBI

1 Introduction

The CHEMDNER task of BioCreative V addresses the automatic identification of chemical entities in medicinal chemistry patents. To participate in this task, we developed an improved version of our IICE framework, trained with the dataset provided by the organization. Our system combines Conditional Random Fields, implemented by Stanford NER [1], brown clustering, implemented by Percy Liang's C-based algorithm [2] and a semantic similarity based on the h-index concept on the ChEBI ontology [3].

This paper provides a description of the system we developed for this task and the results obtained. We submitted 5 runs to the CEMP and CPD subtasks, to test different methodologies that our system can employ.

2 Systems description and methods

To participate in the CHEMDNER task, we adapted the IICE framework, training classifiers with the task dataset. The main part of our system was developed

* Corresponding author: alamurias@lasige.di.fc.ul.pt

in Python 2.7, although it incorporates tools and libraries programmed in other languages. In the context of the competition, we ran the system in batch mode, but it can also run in web service mode [3]. The source code of the framework will be available soon [4]. Running in batch mode, our system was able to classify each document in approximately 3 seconds.

2.1 Pre-processing

Our system performs some basic pre-processing on the input texts, with the objective of facilitating the application of machine learning algorithms. The abstracts are split into sentences using the GENIA sentence splitter [4]. The model used by the splitter was trained with the GENIA corpus, therefore it should perform reasonably well on biomedical texts such as patents. Afterwards, we used the Stanford CoreNLP pipeline [5] to tokenize and extract additional information from the text. Since this pipeline is not tuned for biomedical texts, we added more tokenization rules to split parentheses, dashes and slashes.

2.2 CRF classifiers

The training set of the CEMP subtask was used to train a classifier using the Conditional Random Fields (CRFs) algorithm [6], implemented by Stanford NER. Although classifiers for other types of text are included in the tool, we trained our own, using the SBIEO tagging system and a relatively simple set of features.

To boost the recall of our system, we also trained classifiers for each type of chemical entity included in the annotations. When testing with new text, we merged the results of each classifier, resolving the overlaps in post-processing. Alternatively, we used the results of each classifier as features for a Random Forests (RF) classifier. In this case, we divided the training set into two subsets: the first subset was used to train the CRF classifiers while the results of these classifiers on the second subset were used to train the RF classifier.

2.3 Brown clusters features

Since Stanford NER is able to use features based on word clusters, we experimented with brown clustering to improve the performance of the classification. It has been shown that features based on this information are able to improve the generalization of a classifier when classifying words that did not appear on the training set [7]. For the implementation of the brown clusters, we used Percy Liang's C-based algorithm[5] since it has been previously used for this purpose [8]. The corpus used in the algorithm is composed of abstracts and titles from

[3] http://www.lasige.di.fc.ul.pt/webtools/iice/
[4] https://github.com/AndreLamurias/IBEnt
[5] https://github.com/percyliang/brown-cluster

4 years (2012-2015) of patents from Google's USPTO patents[6]. These patents were selected according to the presence of ChEBI chemical compound names in the title of the patent. The corpus was then pre-processed for the brown clustering algorithm by splitting the text into phrases using, once again, the GENIA sentence splitter, and then, a script that removes the punctuation, to keep the integrity of the chemical compound mentions.

This brown clustering algorithm contains 3 main parameters that can be varied: the number of clusters, the maximum length of a phrase to consider (`plen`) and the minimal occurrence of a phrase (`min-occur`). We conducted our tests by varying the values from 0-1000 clusters, 1-5 `plen` and 1-20 `min-occur`. We found that the best results were obtained using 100 clusters, a `min-occur` of 15 and a `plen` of 2.

2.4 ChEBI mapping and semantic similarity

One of the main features of our system is the integration with the ChEBI ontology. Each entity identified is mapped to a concept, using a lexical similarity method [9] which assigns a confidence score to the mapping. Then, the maximum semantic similarity of each entity to the other entities of the same document is calculated. This was accomplished using a measure, h-simGIC, developed and tested for this particular purpose, based on the concept of h-index, adapted to the ChEBI ontology [3]. The semantic similarity value is used as a measure of confidence of the recognition of a chemical entity. As such, it is possible to assign a threshold, to filter out entities with low semantic similarity.

2.5 Post-processing rules

To further improve the precision of the results, we implemented some post-processing rules, based on the error analysis performed on the development set:

- Discard an entity if contains words from a list of stop words;
- Discard an entity if it does not have any alphabetic characters;
- Remove parentheses that are never closed or opened;
- When two entities identified by different classifiers overlap, choose the longest.

3 Results

We combined the methods employed by our system to run 5 different configurations on the test set. The objective was to compare the results of different approaches and to explore the limits of our system in terms of precision and recall. We set up 3 balanced runs (Runs 1, 2 and 5), one tuned for maximum recall (Run 3) and another tuned for maximum precision (Run 4). The balanced runs consisted in training one CRF classifier with the CEMP training set (Run 2) and training multiple classifiers with half of the training set and then training

[6] http://www.google.com/googlebooks/uspto-patents-grants-text.html

an RF classifier with the results on the other half (Run 1). Run 5 is a variation of Run 2 where we included features based on brown clustering. On Run 3, we used various classifiers trained with the training set and did not apply post processing rules, while on Run 4 we applied a semantic similarity filter of 0.4. This value was determined by testing a range of values on the development set. The results for the CPD subtask were derived from the results for the CEMP subtask: if a document contained an entity identified by our system, it was classified as 1, otherwise, it was classified as 0. The results on Table 1 refer to the precision, recall and F-measure values obtained with each run for the CEMP subtask.

Table 1. Official results obtained with the test set for the CEMP and CPD subtask. Precision (P), Recall (R) and F-measure(F) refer to the CEMP subtask, while Specificity (Spec), Sensitivity (Sens) and Mathew's Correlation Coefficient (MCC) refer to the CPD subtask.

Run	Description	P	R	F	Spec	Sens	MCC
1	RF classifier	79.399%	71.852%	75.437%	88.076%	82.298%	67.559%
2	CRF	85.607%	83.726%	**84.656%**	85.497%	94.142%	**80.254%**
3	Multiple CRF	48.416%	**92.327%**	63.521%	59.429%	**97.875%**	65.99%
4	CRF, similarity >0.4	**86.981%**	42.941%	57.497%	**94.271%**	42.751%	60.157%
5	CRF with clusters	85.803%	83.422%	84.596%	85.603%	93.786%	79.834%

4 Discussion

We were able to achieve our best F-measure with Run 2, by using just one CRF classifier. Adding the brown clustering features did not improve these results. However we can still improve this methodology by using a larger corpus and testing a wide range of clustering parameters. As we expected, Run 4 achieved our highest precision, while Run 3 achieved our highest recall, which was the third highest of this subtask. The results obtained for the CPD subtask were similar, since the highest sensitivity was obtained with run 3, highest specificity with run 4 and highest MCC with run 2. The specificity of Run 2 was the highest that was obtained by any team on the CPD subtask.

To analyze the most common errors made by our system, we chose two chemical entities that were often considered as False Positives (FP) and False Negatives (FN) in the results: "calcium" (15 FPs and 26 FNs) and "alkyl" (58 FPs and 52 FNs). Regarding "alkyl", nearly all errors were made because the system did not identify correctly the boundaries of the entity, which was almost always a chemical entity of the type Family. In the case of "calcium", the FPs were due to instances were the word was not a reference to a chemical entity in that context, for example, in "calcium channels". All FN errors related to calcium were boundary errors. This type of errors could be reduced by using the wisdom of the crowd to create a dataset of corrections and use that dataset to train an improved classifier [10].

Acknowledgments. This work was supported by the Fundação para a Ciência e a Tecnologia (https://www.fct.mctes.pt/) through the PhD grant PD/BD/106083/2015 and LaSIGE Unit Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

References

1. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.
2. Percy Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, 2005.
3. Andre Lamurias, João D Ferreira, and Francisco M Couto. Improving chemical entity recognition through h-index based semantic similarity. *Journal of cheminformatics*, 7(Suppl 1 Text mining for chemistry and the CHEMDNER trac):S13–S13, 2014.
4. Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In *Proceedings of the Second BioCreative Challenge Workshop*, pages 209–212, 2007.
5. Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.
6. John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001*, 2001.
7. Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. Learning distributed word representations for natural logic reasoning. In *Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches: Papers from the 2015 AAAI Spring Symposium*, pages 10–13. AAAI Publications, March 2015.
8. Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.
9. Francisco M Couto, Mário J Silva, and Pedro M Coutinho. Finding genomic ontology terms in text using evidence content. *BMC bioinformatics*, 6(Suppl 1):S21, 2005.
10. Andre Lamurias, Vasco Pedro, Clarke Luka A, and Francisco M Couto. Annotating biomedical ontology terms in electronic health records using crowd-sourcing. In *Proceedings of International Conference on Biomedical Ontology 2015*, 2015.